Inference and Representation, Fall 2015

Problem Set 6: Variational inference

Due: Tuesday, December 8, 2015 at 3pm (uploaded to NYU Classes.)

Your submission should include a PDF file called "solutions.pdf" with your written solutions, separate output files, and all of the code that you wrote.

Important: *See problem set policy on the course web site.*

Latent Dirichlet allocation (LDA) is a probabilistic model for discovering topics in sets of documents [1]. The generative model is as follows:

- For each document, $m = 1, \ldots, M$
  1. Draw topic probabilities $\theta_m \sim p(\theta|\alpha)$
  2. For each of the N words:
     (a) Draw a topic $z_{mn} \sim p(z|\theta_m)$
     (b) Draw a word $w_{mn} \sim p(w|z_{mn}, \beta)$,

where $p(\theta|\alpha)$ is a Dirichlet distribution, and where $p(z|\theta_m)$ and $p(w|z_{mn}, \beta)$ are Multinomial distributions. Treat α and β as fixed hyperparameters. Note that β is a matrix, with one column per topic, and the Multinomial variable z_{mn} selects one of the columns of β to yield multinomial probabilities for w_{mn}.

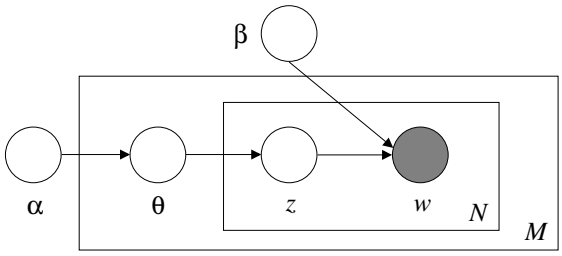

Figure 1: Graphical structure of the LDA model.

1. Derive a mean-field algorithm for inference in the LDA model by minimizing the KL-divergence $D(q_{\gamma_m,\phi_m}(\theta, \mathbf{z})||p(\theta, \mathbf{z} \mid \mathbf{w}))$ with respect to the variational parameters ϕ and γ, where $q_{\gamma,\phi}(\theta, \mathbf{z}) = q_\gamma(\theta) \prod_n q_{\phi_n}(z_n)$, $q_\gamma(\theta)$ is a Dirichlet, and the $q_{\phi_n}(z_n)$ are Multinomial.

   For this question, it is permissible to consult external resources – such as the LDA paper [1] – to help you figure out the derivation (please cite any sources you used). In particular, you will probably want to use the fact that:

$$\mathrm{E}_{q_\gamma}[\log \theta_i] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j).$$

   Important: show all steps of your derivation.

2. Implement the inference algorithm that you derived. You will then run your mean-field algorithm to find the posterior topic distribution θ for an input document, and compare to your results of running Gibbs and collapsed Gibbs sampling from Problem Set 4.

   We have previously learned the parameters (i.e., α and β) of a 200-topic LDA model on a corpus containing thousands of abstracts of papers from the top machine learning conference, Neural Information Processing Systems (NIPS). Your task will be to infer the topic distribution for a new document.

   We have provided the following data files (same as in Problem Set 4):

   - `alphas.txt`, which has on each line for topic i: i, α_i, and a list of the most likely words for this topic,
   - `abstract_*.txt`, with the words of document m (i.e., the abstract),
   - `abstract_*.txt.ready`, with, in order,
     - the number of topics k,
     - α_i, for $i = 1, \ldots, k$,
     - for every word w_n, the word itself followed by $\beta_{w_n,i}$ for $i = 1, \ldots, k$.

   Note that your code only needs to read in the `abstract_*.txt.ready` files – the `alphas.txt` and `abstract_*.txt` files are provided for your reference only.

   For each of the abstracts,

   (a) Plot the ℓ_2 error on your estimate of $E[\theta]$ as a function of the number of iterations for each of the three algorithms (mean-field, Gibbs, collapsed Gibbs). As in Problem Set 4, use $E[\theta]$ computed using collapsed Gibbs sampling with a high number of iterations (e.g. 10^4) as the ground truth.

   (b) Which algorithm converges fastest? Do all algorithms return an accurate estimate of $E[\theta_m]$ when run for a sufficiently long time? Explain your answers.

   Only include in your solutions the plot for the data file `NIPS2008_0517`. The remaining files are provided for your own experimentation.

   You may use the programming language of your choice. We recommend first checking that packages are available to (1) sample from a Dirichlet distribution, and (2) compute the Digamma function $\Psi(x)$, as these will simplify your coding. For example, see Python's `numpy.random.mtrand.dirichlet` and `scipy.special.psi`.

References

[1] David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.